ExtremITA at EVALITA

Evaluation of NLP and Speech Tools for Italian

Multi-Task Sustainable Scaling to Large Language Models at its Extreme

C.D. Hromei, D. Croce

Associazione Italiana di Linguistica Computazionale

UNIVERSITÀ DI TORINO

Università di Roma Tor Vergata

OVERVIEW

- Motivation
- LLaMA: Architecture & Role
- Multi-Task Prompting
- Sustainable training
- Prompting in NERMuD
- Results

Motivation

- Recent development of LLaMA (Touvron et al., 2023) foundational models has opened new ways of exploiting natural language for task specific training through the use of prompting
- Multi task joint training for a single monolithic architecture is appealing when coupled with instructions
- Tasks are modelled as linguistic problems (see ChatGPT): from traditional classification to solve a task to prompting and natural language inference

H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023

EVALITA 2023

https://www.evalita.it/campaigns/evalita-2023/tasks/
13 tasks (22 subtasks)

- **Affect**
  - EMit – Categorical Emotion Detection in Italian Social Media
  - EmotivITA – Dimensional and Multi-dimensional emotion analysis
- **Authorship Analysis**
  - PoliticIT – Political Ideology Detection in Italian Texts
  - GeoLingIt – Geolocation of Linguistic Variation in Italy
  - LangLearn – Language Learning Development

https://www.evalita.it

EVALITA 2023

https://www.evalita.it/campaigns/evalita-2023/tasks/
13 tasks (22 subtasks)

- **Computational Ethics**
  - HaSpeeDe 3 – Political and Religious Hate Speech Detection
  - HODI – Homotransphobia Detection in Italian
  - MULTI-Fake-DetectiVE – MULTImodal Fake News Detection and VErification
  - ACTI – Automatic Conspiracy Theory Identification
- **New Challenges in Long-standing Tasks**
  - NERMuD -Named-Entities Recognition on Multi-Domain Documents
  - CLinkaRT – Linking a Lab Result to its Test Event in the Clinical Domain
  - WiC-ITA – Word-in-Context task for Italian
  - DisCoTEX – Assessing DIScourse COherence in Italian TEXts

https://www.evalita.it

Our Method

22

Istruzione: Quali emozioni sono espresse in questo testo? Puoi scegliere una o più emozioni tra ‘Rabbia’, ‘Anticipazione', ‘Disgusto', ‘Paura', ‘Gioia', ‘Amore', ‘Tristezza', ‘Sorpresa', ‘Fiducia', o ‘Neutro'.
Input: “*Che bella giornata*”

Istruzione: Scrivi le menzioni di entità nel testo, indicandone il tipo: [PER] (persona), [LOC] (luogo), [ORG] (organizzazione).
Input: “*La Banca d’Italia*”

...

Istruzione: Quanto è coerente questa frase, su una scala da 0 a 5?
Input: “*Che bella giornata*”

LLM-based Decoder

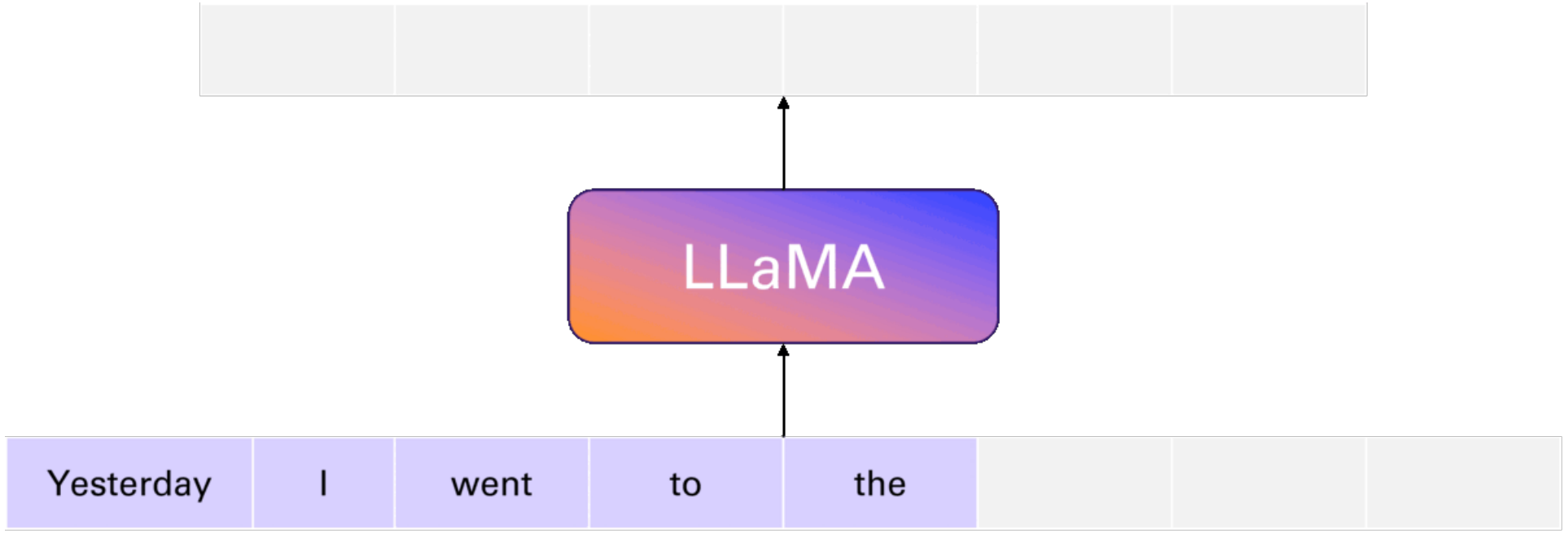
Gioia

[ORG] Banca d’Italia

...

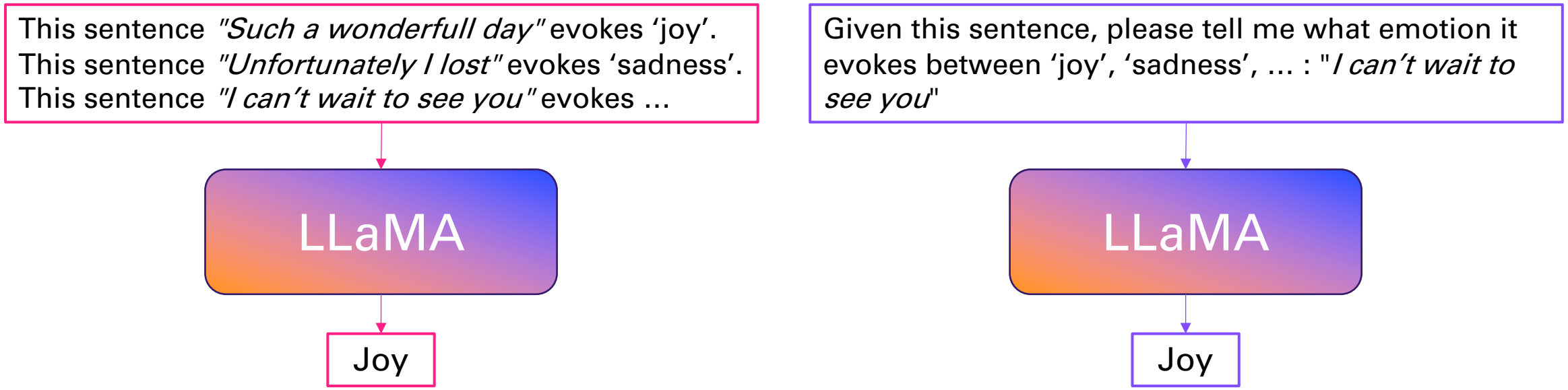
4.0

LLaMA (Touvron et. al 2023): Autoregressive Decoder-only



H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023

LLaMA: Few-shot learning vs Instruction-tuning



There is the need of fine-tuning a multitask architecture (with limited resources) to follow instructions:

- Tasks as Natural Language Inference
- Task independent architecture
- Italian language support
- EVALITA training as Few-Shot Learner

Large Language Models: the Camelids tree

175 Self-Instruct seed tasks

52k Instruction-following examples

Alpaca (Taori et al. 2023)

LLaMA

Example generated task

Instruction:
Brainstorm creative ideas for designing a conference room

Output:
.. Incorporate flexible components, such as..

Instruction: Extract all the entities from the next sentence:
Input: "*Garibaldi , sailing from Caprera, today would know where to go*"

Garibaldi e Caprera

- ✓ Tasks as Natural Language Inference
- ✓ Task independent architecture
- ✗ Italian language support
- ✗ EVALITA training as Few-Shot Learner

R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Stanford Alpaca: An instruction-following llama model, 2023

Large Language Models: the Camelids tree

Text-davinci-003

175 Self-Instruct seed tasks

52k Instruction-following examples

Alpaca

LLaMA

Camoscio(Santilli 2023)

52k Instruction-following examples

EVALITA Data

ExtremITLLaMA

Example EVALITA data

Istruzione:
In questo testo si parla di una cospirazione? Rispondi sì o no.

Input:
Hanno votato tutti obbligo vaccinale, ...

Risposta:
Sì

- ✓ Tasks as Natural Language Inference
- ✓ Task independent architecture
- ✓ Italian language support
- ✓ EVALITA training as Few-Shot Learner

Santilli, A., Rodolà, E.: Camoscio: an Italian instruction-tuned LLaMA 2023

Task Prompts

Task name	Natural language instruction
NERMuD	Scrivi le menzioni di entità nel testo, indicandone il tipo: [PER] (persona), [LOC] (luogo), [ORG] (organizzazione).

Task Prompts

Task name	Natural language instruction
NERMuD	Scrivi le menzioni di entità nel testo, indicandone il tipo: [PER] (persona), [LOC] (luogo), [ORG] (organizzazione).
EMit A	Quali emozioni sono espresse in questo testo? Puoi scegliere una o più emozioni tra 'rabbia', 'anticipazione', 'disgusto', 'paura', 'gioia', 'amore', 'tristezza', 'sorpresa', 'fiducia', o 'neutro'.
EMit B	Di cosa parla il testo, tra 'direzione', 'argomento', 'entrambi', 'non specificato'
EmotivITA	Scrivi quanta valenza è espressa in questo testo su una scala da 1 a 5, seguito da quanto stimolo è espresso in questo testo su una scala da 1 a 5, seguito da quanto controllo è espresso in questo testo su una scala da 1 a 5.
PoliticIT	Scrivi se l'autore del testo è 'uomo' o 'donna', seguito dalla sua appartenenza politica tra 'destra', 'sinistra', 'centrodestra', 'centrosinistra'.
GeoLingIT	Scrivi la regione di appartenenza di chi ha scritto questo testo, seguito dalla latitudine, seguita dalla longitudine.
LangLearn	Questi due testi separati da [SEP] sono presentati nell'ordine in cui sono stati scritti? Rispondi sì o no.
HaSpeeDe 3	In questo testo si esprime odio? Rispondi sì o no.
HODI A	In questo testo si esprime odio omotransfobico? Rispondi sì o no.
HODI B	Con quali parole l'autore del testo precedente esprime odio omotransfobico? Separa le sequenze di parole con [gap].
Multifake-Detective	L'evento riportato nel testo è 'certamente vero', 'probabilmente vero', 'probabilmente falso', o 'certamente falso'?
ACTI A	In questo testo si parla di una cospirazione? Rispondi sì o no.
ACTI B	Di quale teoria cospirazionista parla questo testo, tra 'Covid', 'Qanon', 'Terrapiattista', 'Russia'?
CLinkaRT	Trova i risultati dei test e delle misurazioni nel testo. Per ogni risultato, scrivi '[BREL]', seguito dal risultato seguito da '[SEP]', seguito dal test, seguito da '[EREL]'. Se non trovi nessun risultato, scrivi '[NOREL]'.
WiC-ITA	La parola compresa tra [TGTS] e [TGTE] ha lo stesso significato in entrambe le frasi? Rispondi sì o no.
DisCoTEX 1	Le due frasi precedenti, separate da '[SEP]', sono coerenti tra loro? Rispondi sì o no.
DisCoTEX 2	Quanto è coerente questa frase, su una scala da 0 a 5?

Acceptable Answers

Task name	Output Templates
EMit A	{“Rabbia", “Anticipazione", “Disgusto", “Paura", “Gioia", “Amore", “Tristezza", “Sorpresa", “Fiducia"}+ ∨ “Neutrale"
EMit B	{“Direzione", “Argomento", “Entrambi", “Non specificato"}
EmotivITA	“Valenza: {0-5} Stimolo: {0-5} Controllo: {0-5}"
PoliticIT	“Gender: {“Uomo”, “Donna”} PIB: {“Sinistra”, “Destra”} PIM: {“Sinistra”, “Destra”, “Centro Sinistra”, “Centro Destra”}"
GeoLingIT	“Regione: {Abruzzo, .., Veneto} Latitudine: {} Longitudine: {}"
LangLearn	{“Sì", “No"}
HaSpeeDe 3	{“Sì", “No"}
HODI A	{“Sì", “No"}
HODI B	<Homotransphobia_mention>
Multifake-Detective	{“Certamente Falso", “Probabilmente Falso", “Probabilmente Vero", “Certamente Vero"}
ACTI A	{“Sì", “No"}
ACTI B	{“Terrapiattista", “Covid", “Qanon", “Russia"}
NERMuD	[<entity_type>] <text_span_that_evokes_entity>
CLinkaRT	“[BREL] <Rml_entity_mention> [SEP] <Event_entity_mention> [EREL]"
WiC-ITA	{“Sì", “No"}
DisCoTEX 1	{“Sì", “No"}
DisCoTEX 2	{0-5}

Prompting in NERMuD

- From a token classification problem to a sequence to sequence generation
- Heuristics applied to reconstruct back the original form of the desired output

Istruzione: Scrivi le menzioni di entità nel testo, indicandone il tipo: [PER] (persona), [LOC] (luogo), [ORG] (organizzazione).

Input: *Le dispute tuttavia fra cattolici non cessarono [...] gli argomenti che furono opposti in Germania alla nomina di Marx e [...] collaborazione coi socialisti vennero fatte in Italia.*

[LOC] Germania [PER] Marx [LOC] Italia

Construction of NERMuD data

Original data

Transformation in sequence to sequence

Le	O
dispute	O
tuttavia	O
...	...
in	O
Germania	LOC
alla	O
nomina	O
di	O
Marx	PER
...	...

Input data

Le dispute tuttavia ... in Germania ...

L' ordine n. 14 è senza dubbio una misura ...

Chi è questo Flor ?

[LOC] Germania [PER] Marx [LOC] Italia

nessuna

[PER] Flor

[LOC] Germania [PER] Marx [LOC] Italia

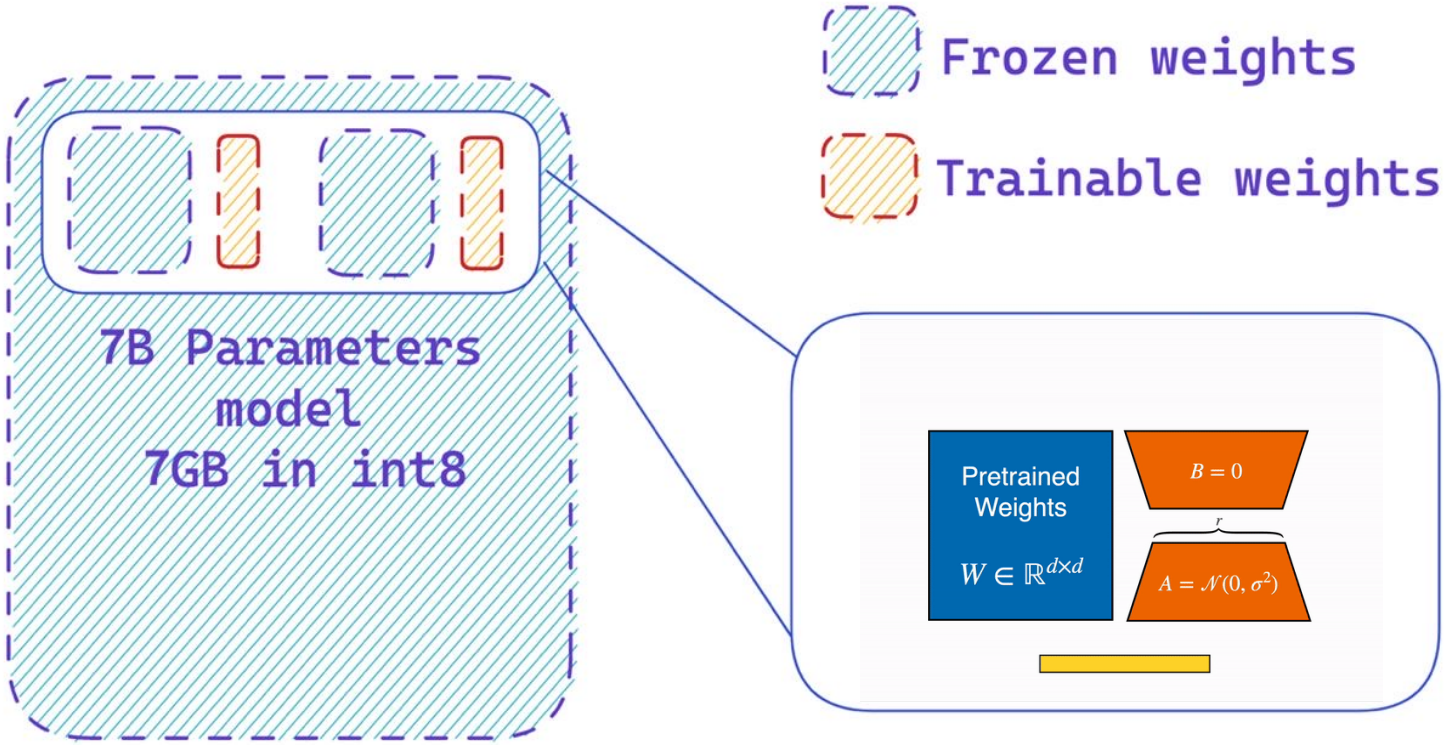
Look up indexes and pair words with labels

Sustainable training

- Low Rank Adaptation (LoRA: Hu et al., 2021): create the parallel (fine-tunable) adapters as smaller matrices
  - add the adapters to the base model while keeping the base model frozen
- We can further scale down the memory required by using int8 approximation, instead of float32

As a result, we successfully fine-tuned a 7B parameters model with a T4 GPU (16GB memory).

Training took 144 hours for 2 epochs on a dataset composed of 134,018 examples.



Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Chen, W.: Lora: Low-rank adaptation of large language models. 2021

Results

Task name	Eval Metric	ExtremITLLaMA Score	ExtremITLLaMA Rank	Best Competitor Score	Best Competitor Rank
EMit	F1	**0.6028** **0.6459**	**1** **1**	0.4994 0.6184	3 3
EmotivITA	Pears Val Pears Aro Pears Dom	**0.8110** **0.6330** **0.6300**	**1**	0.8110 0.6520 0.6540	2
PoliticIT	F1	0.7719	3	**0.8241**	**1**
GeoLingIT	F1 Avg Km	0.3818 145.15	11 9	**0.6630** **97.74**	**1** **1**
LangLearn	F1 F1	0.5500 0.6100	8 8	**0.7500** **0.9300**	**1** **1**
HaSpeeDe 3	F1 – text/context/xPolitic F1 – xRel	0.9034 **0.6525**	3 **1**	**0.9128** 0.6461	**1** 2
HODI	F1 F1	0.7942 **0.7228**	5 **1**	**0.8108** 0.7051	**1** 2
Multifake-Detective	F1 F1	0.5070 **0.4640**	2 **1**	**0.5120** 0.4600	**1** 2
ACTI	F1 F1	0.8565 0.8556	2 5	**0.8571** **0.9123**	**1** **1**
NERMuD	F1	**0.8900**	**1**	na	na
CLinkaRT	F1	0.5916	2	**0.6299**	**1**
WiC-ITA	F1 it-it F1 it-en	0.5100 0.5400	10 8	**0.7300** **0.7400**	**1** **1**
DisCoTEX	Acc HM*	**0.8150** **0.6500**	**1** **1**	0.7200 0.6300	2 2

NERMuD: error analysis

Text	Gold Standard	ExtremITLLaMA
Informa il Consiglio che Nenni gli ha chiesto se il governo porrà ostacoli all' ingresso in Italia di dieci membri dell' Esecutivo internazionale dei partigiani della pace .	[ORG] Consiglio [PER] Nenni [LOC] Italia [ORG] Esecutivo internazionale dei partigiani della pace ✓	[ORG] Consiglio [PER] Nenni [LOC] Italia ✗
Occorre ricordare che l' America vincerà , passerà molto tempo ma vincerà e questa adesione nostra salva il futuro dell' Italia .	[ORG] America [ORG] Italia ✓	[LOC] America [LOC] Italia ✗
Intervento al Senato della Repubblica	[ORG] Senato [ORG] Repubblica ✗	[ORG] Senato della Repubblica ✓
Bisognava capire una buona volta che con un Governo come il nostro , con partiti nemici come abbiamo noi , a dir « tutto o nulla » ci restava e ci resterà sempre la seconda parte .	nessuna ✗	[ORG] Governo ✓
Netflix rimuove un documentario sull' AIDS a seguito di alcune proteste	nessuna ✗	[ORG] Netflix ✓
Russia	[ORG] Russia ?	[LOC] Russia ?

Conclusions

9/22
(41%)

14/22
(64%)

- ✓ One architecture for all tasks
- ✓ Task-independent architecture
- ✓ Sequence to sequence
- ✓ Excellent performance
- ✓ Straightfoward NL prompts
- ✓ Task dependent prompts
- ✓ Code on GitHub and models on HuggingFace

- × Data hungry
- × GPUs and time hungry
- × Non satisfactory performance on all tasks:
  - × GeoLingIT
  - × LangLearn
  - × WiC-ITA

Now it is your turn

Lab Objectives: Instruction Tuning with Q-LoRA

Goal:

- Hands-on lab to explore **instruction tuning** of LLaMA using **Q-LoRA** on a real task from EVALITA 2023.

What you'll do:

- Format a task as prompt-based input/output
- Fine-tune a quantized LLaMA (7B) using Q-LoRA
- Evaluate performance with supervised metrics
- Reflect on benefits vs prompting-only methods

Resources & Setup

The lab is based on the BISS 2024 GitHub repository:

https://github.com/crux82/BISS-2024

Focus on the Lab2/ folder: fine-tuning a LLaMA-based model using Q-LoRA on EVALITA tasks.

Recommended environment: Google Colab with T4 GPU (16GB)

Colab notebooks:

- Prompt encoding
  - https://colab.research.google.com/github/crux82/BISS-2024/blob/main/Lab2/1_prompt_encoding.ipynb
- Q-LoRA fine-tuning
  - https://colab.research.google.com/github/crux82/BISS-2024/blob/main/Lab2/2_train_lora.ipynb
- Inference
  - https://colab.research.google.com/github/crux82/BISS-2024/blob/main/Lab2/3_inference.ipynb
- Decoding and evaluation
  - https://colab.research.google.com/github/crux82/BISS-2024/blob/main/Lab2/4_decode_and_evaluate.ipynb

Suggested Exercises

- Fine-tune the model on **CLinkaRT** (clinical causal relations)
- Compare Q-LoRA results with prompting-only approach (seen in last lab)
- Modify the prompt style and observe effects on generalization
- Evaluate performance using exact match / custom metrics